*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

**Big Data y Machine Learning para Economía Aplicada**

**Taller 2**

**2023-02**

**Link del repositorio en Github:** [**https://github.com/djfarfan10/Taller-2**](https://github.com/djfarfan10/Taller-2)

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Angie Ariza Quitian - 201325848

**Presentado por**

Andrés Felipe Diaz Barreto - 200610686

Duvan Javier Farfán López – 201317299

Daniel Eduardo Sandoval - 200712968

• Introducción. La introducción expone brevemente el problema y si existen antecedentes. Describe brevemente los datos y su idoneidad para abordar la pregunta del conjunto de problemas. Contiene un avance de los resultados y las conclusiones principales.

• Datos. En este conjunto de problemas, se requiere que amplíes las variables en tus datos (recuerda ampliar los datos de entrenamiento y prueba). Como mínimo, debes agregar seis variables adicionales:

De acuerdo con la literatura relacionada con los precios de las viviendas, este se suele determinar a través del modelo de precios hedónicos, que busca estimar el valor de un bien inmueble a través de las características que poseen los mismos y que los hacen diferenciados de otros (Taylor, L. 200X). Es así como surgen variables o características como la cercanía a lugares de servicios (como hospitales, colegios, parques, estaciones de policía, entre otros), y la estructura propia de la construcción que pueden influir en el valor de los inmuebles. Estás características se basan en aquellos atributos propios de los apartamentos y casas del análisis, así como de la ubicación. Según la literatura, dentro de las características propias de un inmueble, los atributos que se destacan son la cantidad de habitaciones, dormitorios y baños que posee, así como la antigüedad de los mismos, mientras que en cuanto a la ubicación se relacionan aspectos de bienes no mercadeables como la calidad del aire, el ruido, la seguridad, entre otros.

Con el objetivo de predecir los precios de viviendas en Chapinero se usaron dos bases de datos, *train* y *test*, con información asociada a casas y apartamentos en Bogotá, las cuales contenían 38.644 y 10.286 observaciones respectivamente. Al realizar una primera inspección de cada una de las bases se observó que algunas variables importantes para la estimación de precios como el área de los inmuebles, el número de baños y habitaciones contaba con una gran cantidad de missing values. También se observo que de la base de *train* no se contaba con la descripción de 9 observaciones, por lo que se eliminaron al no ser de utilidad para la extracción de la información faltante.

Teniendo en cuenta lo anterior, se extrajo información a partir de la columna *description* para áreas y baños, y se comparo con la información que ya se suministraba. Para cada caso se seleccionó el mayor número de la observación para dichas variables y se consolidó la información en una columna final. Dado que no fue posible extraer la totalidad de información, para los valores de NAN de las variables de área y baños se imputaron estas observaciones faltantes por la mediana de cada variable.

Se realizó también la extracción del tipo de vivienda de la descripción y se comparo con la suministrada en la base, defendiendo como la información final la extraída del texto suministrado en cada base de datos.

Además de las variables previamente mencionadas, usando herramientas de código para extracción de texto se crearon nuevas variables que permitieron identificar las características propias de cada vivienda en la muestra. Las variables creadas se pueden ver en la tabla 1.

Tabla 1.

|  |  |
| --- | --- |
| Variables adicionales extraídas del texto | Descripción |
| Parqueadero | Variable dicótoma: 1 si la vivienda tiene parqueadero, 0 si no tiene |
| Baño social | Variable dicótoma calculada a partir de información de extraída para baños y habitaciones: 1 si la cantidad de baños es mayor a el numero de habitaciones, 0 si no lo es |
| Deposito | Variable dicótoma: 1 si la vivienda cuenta con deposito o garaje, 0 si no tiene |
| Estado construcción | Variable dicótoma: 1 si la vivienda es nueva, 0 si no lo es |
| Estado remodelado | Variable dicótoma: 1 si la vivienda está remodelada, 0 si no lo es |
| Terraza o balcón | Variable dicótoma: 1 si la vivienda tiene terraza o balcón, 0 si no cuenta con ninguno de los dos |

Fuente: Elaboración propia

Los valores que no se pudieron extraer (NAN) para las variables de la tabla 1 fueron reemplazados con 0, con la finalidad de no sobreestimar los precios de las viviendas, teniendo en cuenta el objetivo del problema set.

**Variables externas**

En esta sección se describen las variables que fueron tomadas de fuentes externas a las bases de datos proveídas en el ejercicio. La selección de alguna de estas variables se hizo a partir de un estudio anterior (Toloza-Delgado et al., 2021), en el que identifican la importancia que tienen el estrato socioeconómico, la distancia a estaciones de TransMilenio y la distancia a parques.

Por otro lado, de acuerdo con Chacón Rodríguez, V. 2018), los instrumentos de Unidad de Planeamiento Zonal (UPZ) –reemplazadas en el 2022 por las Unidades de Planeamiento Local (UPL)– fueron utilizados para agrupar zonas homogéneas de la ciudad, por lo que pueden ser segmentaciones que capturen parte del precio de la vivienda.

Distancia a estaciones de TransMilenio

Las ubicaciones de las estaciones de TransMilenio fueron tomadas de la página principal de la entidad (disponibles en un archivo geojson), teniendo en cuenta que al utilizar el atributo de paradas de buses en la API de OpenStreetMap salían mezcladas algunas cuantas estaciones para buses que no son BRT. En total, se tienen 149 estaciones de TransMilenio distribuidas en 12 troncales del sistema.

Estratos socioeconómicos

En la plataforma distrital de Datos Abiertos, se tiene, entre otros, las manzanas georreferenciadas según su estrato socioeconómico. Este atributo es asignado por las alcaldías para clasificar a una vivienda según características físicas externas e internas, su entorno inmediato y su contexto habitacional o funcional (DANE, 2023).

Por la georreferenciación inadecuada de algunas viviendas y porque hay áreas de la ciudad que no están incluidas dentro de estas manzanas (vías, áreas públicas, zonas comerciales), se asignó el estrato de la vivienda según su cercanía al centroide del polígono , es decir, la vivienda tenía el valor del estrato del polígono cuyo centroide estuviera más cerca. En la siguiente figura se muestra el resultado de la asignación:

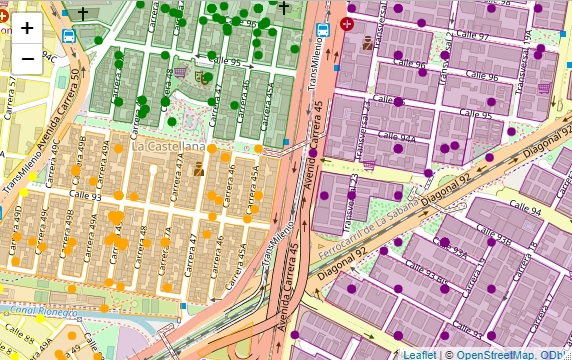


Figura 1. Muestra de asignación de los estratos socioeconómicos a las viviendas. Los puntos señalados en los recuadros rojos son ejemplos de la georreferenciación inadecuada, lo que resulta en que haya viviendas por fuera de las manzanas residenciales.

La zona de la anterior figura es muy interesante porque en un radio de 1000 metros se pueden ver cuatro estratos socioeconómicos: tres (amarillo), cuatro (naranja), cinco (verde) y seis (purpura). Los puntos representan las viviendas de las bases de datos y los polígonos sombreados son los que estaban disponibles en Datos Abiertos

Unidades de Planeamiento Zonal

Los polígonos de la UPZ fueron tomados del Laboratorio Urbano Bogotá (Laboratorio Urbano, 2012) y emparejados con las viviendas de las bases de datos según su ubicación espacial. En total, se tienen 112 UPZ correspondientes 19 localidades dentro de la ciudad.

Así mismo, revisando las categorías contenidas en *leisure* y *amenity* de Open Street Maps, y de acuerdo a la literatura revisada, se obtuvo información de parques, hospitales, universidades, colegios, comandos de atención inmediata (CAIs), con la cual se determinó la distancia mínima de cada inmueble a cada uno de estos espacios. En la figura XX se observa la distribución de las categorías usadas a lo largo de Bogotá.

– Al menos 4 predictores procedentes de fuentes externas; estos pueden ser de mapas de Open Street.

– Al menos 2 predictores procedentes del título o la descripción de las propiedades.

Cuando escribas esta sección, debes:

Describir los datos, su idoneidad para el problema y el proceso de construcción de la muestra, incluyendo cómo se limpiaron los datos, se combinaron y se crearon nuevas variables.

Incluir un análisis descriptivo de los datos. Como mínimo, debes incluir una tabla de estadísticas descriptivas y dos mapas con su interpretación. Sin embargo, se espera un análisis profundo que ayude al lector a comprender los datos, su variación y la justificación de tus elecciones de datos. Utiliza tu conocimiento profesional para agregar valor a esta sección. No lo presentes como una "lista seca" de ingredientes.

• Modelo y Resultados. Esta sección presenta el modelo con la mejor puntuación enviado para su evaluación. Cuando escribas esta sección, incluye:

– Una explicación de las variables utilizadas para entrenar este modelo, recuerda utilizar las variables que agregaste en la sección anterior.

– Una explicación detallada de cómo se entrenó, la selección de hiperparámetros y cualquier otra información relevante.

De acuerdo con la literatura académica relacional con los precios de las viviendas, este se suele determinar a través del modelo de precios hedónicos, que busca estimar el valor de un bien inmueble a través de las características que poseen bienes que diferenciados (Taylor, L. 200X). Es así que, surgen variables o características de los inmuebles como la cercanía equipamientos (hospitales, colegios, parques, estaciones de policía, etc.), así como el área de la propiedad y otra serie de características (ubicación).

Para este ejercicio se determinaron como variables internas a las características de las viviendas, las siguientes variables: tipo de la vivienda (casa o apartamento), el número de habitaciones (dormitorios), si cuenta o no con parqueadero/garaje, el área (tamaño) de la propiedad, el número de baños y si cuenta o no con deposito.

**Tipo de vivienda**

Se refiere a si es una casa o un apartamento. Bogotá es una de las ciudades con mayor densidad en América Latina, por lo que el mercado inmobiliario ha venido cambiando en el que la construcción de apartamentos predomina desde hace varios años.

**Habitaciones**

– Una sección que compare el rendimiento de la presentación con la mejor puntuación con otras presentaciones enviadas a Kaggle. Estas presentaciones deben incluir especificaciones entrenadas utilizando modelos de Regresión Lineal, Ridge, Lasso, Elastic Net, CART, Random Forest y Boosting. Indica en tu archivo de presentación el nombre del modelo utilizado.

• Conclusiones y recomendaciones. En esta sección, expones brevemente las conclusiones principales de tu trabajo.